This article was reprinted from the April 1992 issue of *Information Systems Developments*.

Wide Area Information Servers (WAIS)

By Eliot J. Christian and Timothy L. Gauslin, ISD, Reston

Wide Area Information Servers (WAIS, pronounced "ways") address a fundamental problem: How can you find and retrieve the data and information you need when you don't know what is available, where to find the information, or how to access it?

There are many information services and data base products to help users find and retrieve information, but most are targeted to specialists or require the user to understand how the data is organized. Because most researchers gather information from a variety of sources, their frustrations in dealing with complex and divergent approaches may be inhibiting growth in the information services market. Dow Jones News Service and Peat Marwick teamed with Apple Computer and Thinking Machines to develop a facility whereby users with limited computer skills could access personal, corporate, or published information from one interface—WAIS.

WAIS is developing in the context of an explosive growth of digital information, especially as the definition of information expands beyond alphanumeric to graphics and multimedia (e.g, audio, music, video). Yet, network connectivity through utilities such as the Internet is also growing explosively, as is the availability of computing power ranging from ubiquitous desktop workstations to massively parallel supercomputers. Although conventional brute force methods are capable of searching sources containing several trillion characters of text and data, better techniques are needed to make such searches quick and affordable.

An Open Standard

In the late 1980's, the information services community was completing a crucial piece of work—a standard known as the Information Retrieval Service Definition and Protocol Specification (Z39.50). The importance of open standards in a product such as WAIS cannot be overemphasized. Because a single computer-to-computer protocol is used, information can be stored anywhere on different types of machines, and a single interface can retrieve information from all. As the standard gains wide acceptance, proprietary access mechanisms will give way to competition on the content of information sources while allowing add-on features for special interest communities.

The Z39.50 standard is a product of the National Information Standards Organization (NISO), accredited to the American National Standards Institute. NISO Z39.50 is fully compatible with the NISO standard for library catalogs (Z39.2) originally promulgated by the Library of Congress and known as MARC (Machine Readable Cataloging). Both Z39.50 and Z39.2 have corresponding International Standards Organization (ISO) standards. Z39.50 is an applications service within the family of ISO standards comprising the Open Systems Interconnection model. It is also fully compliant with TCP/IP as implemented on the international network of networks known as the Internet.

Adoption of the Z39.50 standard protocol throughout the international library community fits in well with the WAIS goal of making information searches coherent across different services. WAIS implemented the 1988 version of the Z39.50 protocol with extensions that are now reflected in the 1992 version of Z39.50. The 1992 version, currently in formal review, has a built-in mechanism to assure that implementations interoperate exactly as the standard requires. Computer-to-computer interchanges, whether components of the Z39.50 protocol or of the content being delivered, are precisely represented in a standard computer language known as Abstract Syntax Notation.

Client Server Interaction

WAIS implements Z39.50 in a client/server mode of computer interaction. In a typical search for textual information, the client software prompts the user to select which information sources to include in the search and to enter a search request. Once the search request is entered by the user, the client software converts the search words to the standard information retrieval protocol (Z39.50) and presents the search request in turn to each server associated with a selected source. The server software takes the words and matches them to the contents of all documents in each selected source. The client software receives search results from all of the servers and presents to the user a list

This article was reprinted from the April 1992 issue of *Information Systems Developments.*

of all document titles found. When requested by the user, the client software requests the server to pass the full contents of the document, and the client then presents the document to the user.

Document Scoring

Up to this point, the Z39.50 client/server interaction is fairly conventional, and many information service providers will fit their current products into this model. One very useful feature of Z39.50 is that the user sees documents listed in a ranked order based on relative scores assigned to documents by the servers. The algorithm employed for the scoring of documents will be a fascinating area of development. In a sense, the scoring algorithm is where judgment is applied about the likelihood that a particular document will be seen as relevant by the user. The public domain version of WAIS has a fairly simplistic scoring algorithm which stresses the frequency of occurrence of the searched for words. It does include differential weighing for where in the document the words occur and whether the words occur so often as to be non-specific.

Requests in English

A critical issue in searching for information is that the results of the search can be compromised if the user lacks the skills needed to use the client software to specify the search. Because the WAIS goal is to totally avoid the usual kinds of query languages, a search request in WAIS is simply entered by the user in English. The WAIS software does not analyze grammatical structures or the meanings of the words. It simply depends on the fact that the scoring of documents will tend to emphasize words, primarily nouns, that are more specific. Conversational embellishments will tend to get low scores since they are unlikely to be of specific value in the target documents. While the simplicity of the WAIS search request has a certain appeal, real users cannot be expected to bring with them all of the key words needed to find everything that may be relevant. WAIS therefore provides an elegant mechanism to refine the search so that the session converges on meaningful results—an approach known as "relevance feedback."

Relevance Feedback

A WAIS information search and retrieval session evokes the experience of using a library. A library user may begin by consulting a card catalog or index or by asking a reference librarian for help. At this point, the user is searching for documents based on a few key words (e.g., subject, title) or names (e.g., author). As the user reviews the documents found, he or she may note other key words or names that could lead to additional relevant documents. A feedback situation develops as the user modifies subsequent searches based on results found in prior searches. Ideally, the user stops searching when all the most relevant documents are found.

In WAIS, any retrieved document or part of a document that seems to be getting closer to what the user is interested in may be highlighted. When the user designates that highlighted portion as relevance feedback, WAIS treats the words in that portion as another search request. This ability to have the user refine the search iteratively through relevance feedback is being addressed in the 1992 version of Z39.50. Some WAIS developers see relevance feedback as crucial to achieving the goal of eliminating complicated query languages.

WAIS Servers

WAIS information servers are registered to the Directory of Servers currently maintained on the Internet by Thinking Machines. Registration requires a commitment to keep the server in a reliable operational mode. The registration entry includes text information about the contents of the sources reachable through the server—this information is itself indexed for searching. (Expected in the 1992 version of Z39.50 is guidance on standardizing the mechanism that allows clients to learn about the information sources on a server.) Also listed is information that will be used by the client software to contact the server (e.g., TCP/IP node name), as well as information on what and how to pay charges for use of the server if it is not free.

Any server capable of responding to Z39.50 information retrieval requests can be an information server. Information servers can be local (on the workstation or local area network) as well as remote (accessible now via TCP/IP and X.25 networks or asynchronous dial-up access in the future). WAIS does not require any central coordination unless the server is to be advertised through the Directory of Servers. In fact, an information server registered to the Directory of Servers can itself act as a subordinate directory of servers administered locally. By describing sources under various directories of servers, it is possible to organize the sources in whatever relationships make sense and yet allow users to search as many sources as desired.

One feature of WAIS that is allowed but not required by the Z39.50-1988 protocol is that the client/server interaction is "stateless." At the application level, each request from the client to the server is a separate process that is not associated to any previous request. The server does not maintain information about the client between requests. (For efficiency, the communications link itself may be retained across requests.) This feature is significant for situations in which a user

This article was reprinted from the April 1992 issue of *Information Systems Developments*.

may want to search hundreds of sources on dozens of servers at a sitting.

Information Sources

Information servers exist to provide access to the information sources placed on them. These sources are compilations that may include a variety of formats. Such formats are known as "document types," although information need not be textual. While all Z39.50 clients and servers support search and retrieval of textual information, support for other document types that may have been registered in Z39.50 (e.g., MARC bibliographic format, graphics, hypertext, video) is negotiated when the client initiates its relationship with a server.

When sources are created, defining the document types allows the server to use the appropriate translation between the specific query format of the source and the Z39.50 protocol. The public domain WAIS package includes assistance in creating information sources. Indexing software is provided in the WAIS package for several common document types consisting of text, graphics, and bibliographic references in MARC. Source code in the C programming language is provided for adding other document types. If access to other data structures is required, the server interface routines are also designed to be customized. A typical customization would be to use search requests to access a relational data base such as Ingress or Oracle using Structured Query Language (SQL).

Existing and Proposed WAIS Implementations

There are hundreds of information servers registered in the Directory of Servers. The following list illustrates the variety of uses to which WAIS is being applied. At present, there is no charge for using these servers.

- Thinking Machines maintains the Directory of Servers, WAIS documentation and source code, Frequently Asked Questions, some patents, molecular biology abstracts, a cookbook, the Central Intelligence Agency's *World Factbook*, and weather maps and forecasts.
- Massachusetts Institute of Technology maintains a server of classical and modern poetry.
- Cosmic maintains descriptions of government software packages.
- Library of Congress is creating a WAIS server for their card catalog.
- Columbia Law Library maintains abstracts of legal decisions.
- National Institutes of Health publishes announcements of opportunities for research grants.
- The High Performance Computing and Communications Initiative is looking at applying WAIS.
- Apple has created a dynamic version of WAIS that periodically retrieves requested information in an unattended mode and presents to the user a "personalized newspaper" showing the latest information as it is published.

The range of subject matter that can be covered by WAIS could be viewed as a blessing or a curse depending on whether the user needs a very broad or very narrow search. As in any publishing medium, there is also likely to be a wide range in the credibility of information across sources. WAIS developers are considering whether sources might carry endorsements that they would earn through peer review, in the manner of respected science journals.

WAIS Applications in the USGS

USGS involvement with WAIS stems from the efforts of the Information Systems Division (ISD) to enhance the current version of the Earth Science Data Directory (ESDD). ESDD is maintained as a useful source for references to earth science data, including many at the state level, and a comprehensive list of data holdings relevant for arctic research. WAIS is especially appropriate for that application, because the ESDD user community ranges from local citizenry to international global change researchers. The ability of WAIS to place the ESDD in the context of other USGS and external information sources would be especially powerful for these users. One idea being considered is to publish in the WAIS Directory of Servers a subordinate directory of servers focused on earth science data and information. That directory would be maintained by the USGS but could list sources of other organizations as well.

ISD is adding features required for ESDD (phrase searching, location searching, and key word searching within fields), which can be accommodated within WAIS and the Z39.50 standard. ISD is also including the ability for a user of the client software to drop from a WAIS session into an automated login to existing data systems such as the Global Land Information System. With this approach, users of the USGS/WAIS client software would be able to access any Z39.50 server but would have additional capabilities when accessing one of the USGS servers. A subgroup of the USGS Geographic Data Committee is also exploring the possibility of using WAIS to access a clearinghouse of USGS spatial data holdings.

The following examples represent instances where USGS data and information are already being made available in a digital form. From a technical viewpoint, it would be a small step to put that information onto a WAIS server. Note that until user authentication is included in Z39.50 (it is in the 1992 version but not in the 1988 ver-

This article was reprinted from the April 1992 issue of *Information Systems Developments*.

sion), information servers should not include restricted information.

- Announcements—USGS News Releases, USGS as mentioned in the news, schedule of upcoming events, calls for papers, training opportunity announcements, computer virus alerts, Requests for Proposals.
- Catalogs and publications—USGS Library Catalog, USGS publications including Circulars and Open File reports, conference proceedings, abstracts of publications, newsletters, technology assessment reports, Earth Science Information Center product descriptions and ordering information.
- Data accessed as information—The fully qualified names of data fields can be made to be searchable so that a user could have the equivalent of a Model 204 or SQL query without having to learn the query language.
- Education—The simplicity of the WAIS user interface may make it a good tool for teaching students how to navigate the world of scientific information.
- Electronic filing—Personal, branch-wide, or office-wide files of outgoing correspondence can be maintained using WAIS.
- Full-text search interface for CD-ROM's—WAIS may provide a good base for a standard user interface for text searching which is being pursued by the community of CD-ROM developers and users.
- Graphics—Scientific visualization product demonstrations; architectural drawings such as building and floor layouts of USGS occupied space; forms distribution for local printing, with or without fill-in software.
- Information access from data—Textual and graphic information can be linked to specific data fields in a data base (e.g., station history linked from the data values for that station).
- Management information—Strategic plans, all-employee letters, security guidelines.
- Manuals and handbooks—Hazardous materials information for laboratories, users' manuals such as mainframe users' manual or help files, technical manuals, procedures manuals (e.g., quality assurance procedures), *Survey Manual*, guidelines for ADP acquisition with sample requirements definition documents.
- Personnel—Locator information such as phone directory, organizational directory, electronic mail directory; model position descriptions and performance standards; vacancy announcements; personnel regulations.
- Software—Directories of public domain and shareware, including searching of program descriptions and source code; distribution of software.

This list only scratches the surface of possible applications, and WAIS may well not be the most appropriate vehicle for all of these. There is also a set of issues to be addressed concerning the informal release of digital data and information that might be misperceived as having been through the formal USGS review process.

WAIS and Global Change

WAIS is being used for the enhanced version of the ESDD which interfaces with the Interagency Global Change Master Directory, a single source for references to key global change data. The global change data management community is considering WAIS to rapidly correlate the Global Change Master Directory to other already existing data directories relevant to global change research. For example, the National Oceanic and Atmospheric Administration has a directory with about 25,000 data set references, and the Inter-University Consortium for Political and Social Research has another directory referencing about 28,000 data sets. This approach would be especially useful for the very broad community of users intended to be served by the Consortium for International Earth Science Information Network (CIESIN). WAIS has also been suggested for the "Directory of Directories" effort taking place under the International Council of Scientific Unions' Committee on Data.

The ability of WAIS to handle different information sources through a single user interface makes it possible for researchers to explore publications and data sets concurrently. The federal research libraries involved in global change research (primarily, National Aeronautics and Space Administration, National Oceanic and Atmospheric Administration, U.S. Geological Survey, and U.S. Department of Agriculture) are very interested in the potential for WAIS to bridge between the data and information worlds. Also, WAIS is seen as a useful way to connect textual information into a data system. For example, when a user is researching an existing data set, it would be useful to provide immediate access to all of the associated documentation about that data set. Using WAIS, the associated documentation could extend beyond the data set itself to include publications which reference the data set or engineering specifications of the instruments used.

Because WAIS simplifies searching for information across many dimensions, it would be a natural tool for use in the Global Change Research Information Office. This new office is man-

This article was reprinted from the April 1992 issue of *Information Systems Developments.*

dated by the Global Change Research Act of 1990 to provide information internationally on global change research and technologies associated with mitigating or adapting to global change.

Software Requirements and Support

The public domain WAIS software is written in C and has been ported to a variety of platforms. The client software is available for a range of UNIX workstations (including Sun and Data General) in the X Window System, MS-DOS in either Windows 3.0 (using Dynamic Link Libraries) or character mode, the Apple Macintosh, and the NeXT computers. The server software runs on these computers, plus DEC VMS, IBM VM, and the Connection Machine, among others.

WAIS clients and servers have been implemented without communications on UNIX, Macintosh, and MS-DOS. This allows WAIS to be used also to access data and information distributed on floppy disks or CD-ROM's. Terminals having only asynchronous dial-up capabilities can access WAIS by telephone connection to a client running on a UNIX machine. Any WAIS client can access WAIS servers through TCP/IP software. TCP/IP is usually included with UNIX workstations, and both MS-DOS versions (Windows and non-Windows) include public domain TCP/IP software.

Indexing of text to create an information source is fairly rapid: A 30 megabyte file was indexed in about 20 minutes on a Data General AViiON. In all instances explored so far, searching sources either locally or on the Internet occurs in a matter of seconds.

WAIS Support

In pursuing WAIS in the USGS, ISD will be working closely with the broader WAIS community, which includes about 150 universities and a number of major corporations (Apple, Sun, NeXT, DEC, Microsoft, Dow Jones, Peat Marwick, Mead Data Central, and Thinking Machines, among others). The overall leader of the project thus far has been Thinking Machines, Inc. The University of North Carolina has proposed to the National Science Foundation to become the focus for long-term support of WAIS in a manner similar to the public domain Kermit software support provided by Columbia University. Those heavily involved in WAIS development are working very closely with NISO as well.

Conclusion

WAIS represents more than just another slick new piece of software. Although it is based on the Z39.50 protocol born in the library science and information services community, WAIS bridges to the computer science and data processing services community. Such bridges hold out the promise of revolutionary improvements in information services and in how information is handled in data processing systems.

The capabilities demonstrated by the Z39.50 standard and the WAIS implementation challenge us in the USGS to examine how we think of access to data and information and how we make public USGS data and information.

Further Information

Within the USGS, Tim Gauslin of ISD is currently the focal point for WAIS development and applications and can provide some information on the NISO Z39.50 standard. Because Tim personally ported the WAIS software to MS-DOS and the Data General, he should be contacted for those versions. Tim should also be consulted if there is interest in bringing up USGS information servers and sources in addition to ESDD. Tim may be contacted at:

U.S. Geological Survey
802 National Center
Reston, Virginia
703-648-5980
tgauslin@isdres.er.usgs.gov

Information on WAIS from Thinking Machines is available from Barbara Lincoln (TCP/IP: barbara@think.com), as well as Brewster Kahle of Thinking Machines (TCP/IP: brewster@think.com). ■